Human Interaction and Emerging Technologies (IHIET 2023), Vol. 111, 2023, 194–199
https://doi.org/10.54941/ahfe1004024

Lip-Reading Research Based on ShuffleNet and Attention-GRU

Yuanyao Lu and Yixian Fu

School of Information Science and Technology, North China University of Technology, Beijing, 100144, China

ABSTRACT

Human-computer interaction has seen a paradigm shift from textual or display-based control towards more intuitive control such as voice, gesture and mimicry. Particularly, speech recognition has attracted a lot of attention because it is the most prominent mode of communication. However, performance of speech recognition systems varies significantly according to sources of background noise, types of talkers and listener's hearing ability. Therefore, lip recognition technology which detects spoken words by tracking speaker's lip movements comes into being. It provides an alternative way for scenes with high background noise and people with hearing impaired problems. Also, lip reading technology has widespread application in public safety analysis, animation lip synthesis, identity authentication and other fields. Traditionally, most work in lipreading was based on hand-engineered features, that were usually modeled by HMM-based pipeline. Recently, deep learning methods are deployed either for extracting "deep" features or for building end-to-end architectures. In this paper, we propose a neural network architecture combining convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) with a plug-in attention mechanism. The model consists of five parts: (1). Input: We employ the Dlib library to detect 68 facial landmarks, allowing us to isolate the lip area and extract a series of 29 consecutive frames from the video sequence. These frames are then processed through a basic C3D network to facilitate generic feature extraction. (2). CNN: As neural networks continue to expand in depth, computational complexity rises substantially, leading to the development of lightweight model architectures. Our method employs a lightweight CNN called ShuffleNet, which is pretrained on the ImageNet dataset, to perform spatial downsampling on individual images. ShuffleNet primarily utilizes two innovative operations—pointwise group convolution and channel shuffle—significantly reducing computational costs without compromising recognition accuracy. (3) CBAM: In image processing, feature maps hold a wealth of essential information. Traditional convolutional neural networks apply convolution uniformly across all channels, despite the varying significance of information across different channels. To enhance the performance of convolutional neural networks in feature extraction, we incorporate an attention mechanism called Convolutional Block Attention Module (CBAM). CBAM is a straightforward yet efficient attention module designed for feedforward convolutional neural networks. It consists of two independent sub-modules, Channel Attention Module (CAM) and Spatial Attention Module (SAM), which execute channel and spatial attention, respectively. (4) RNN: T Traditional Recurrent Neural Networks (RNNs) are primarily employed for processing sequential data; however, as RNN networks expand, they may struggle to maintain connections to all relevant information, potentially resulting in the loss of crucial details. This limitation prevents traditional RNNs from addressing long-distance dependency issues, causing a significant drop in performance. To overcome this shortcoming, we opt for the GRU network in our research, a variation of LSTM that offers a simpler structure and superior performance compared to the LSTM neural network. (5) Outputs: Finally, we feed output of the backend to the SoftMax function for final word classification. Through our experimentation, we compare multiple model architectures and discover that our model attains an accuracy comparable to the current state-of-the-art model while requiring less computational resources.

Keywords: Lip reading, CBAM, Light-weight network

INTRODUCTION

Lip reading technology, also known as visual speech recognition or speech reading, is an innovative field that focuses on interpreting human speech through the analysis of lip movements and facial expressions. This technology has advanced significantly in recent years, largely due to breakthroughs in artificial intelligence (AI), computer vision, and deep learning algorithms. The primary goal of lip- reading technology is to bridge the communication gap for individuals with hearing impairments, enabling them to engage more effectively in conversations and social interactions. Additionally, this technology can be utilized in noisy environments, where audio-based speech recognition may struggle to perform accurately. Furthermore, lip reading technology has the potential to enhance security measures, assist in voice-driven device control, and improve teleconferencing experiences. The lip-reading technology has developed through key stages leveraging artificial intelligence, computer vision, and machine learning. First, a diverse dataset of video recordings is collected, capturing various languages, accents, and environments. Next, data preprocessing isolates the relevant lip and facial features. Then, machine learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are employed to train models for recognizing and deciphering lip movements. As these models are continuously trained and refined with new data, their performance improves. Finally, integration with applications and devices allows the technology to assist in real-world communication and accessibility scenarios.

LIP-READING MODEL ARCHITECTURE

The lip-reading technique primarily features two interconnected components: a frontend network and a backend network. The frontend networks comprise a variety of architectures, such as MobileNet, ShuffleNet, VGGNet, GoogLeNet, and ResNet, while the backend networks encompass Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. An attention mechanism is employed to optimize resource allocation by prioritizing crucial information in the feature map. Our proposed system integrates ShuffleNet, CBAM, and GRU, with the architecture depicted in Figure 1.

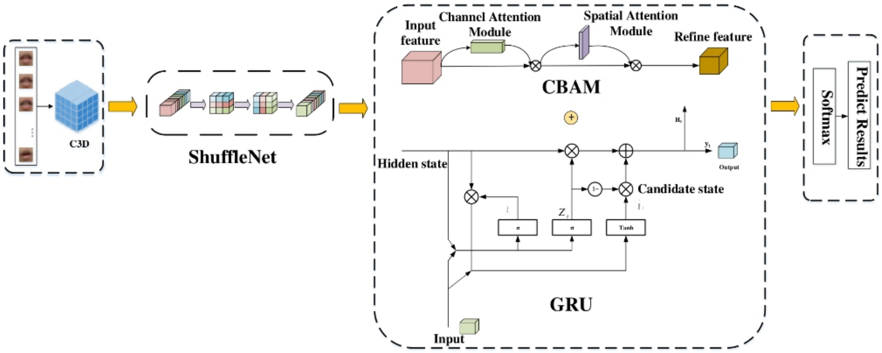

Figure 1: Architecture of the lip-reading model.

The frontend network is ShuffleNet. It is an optimized convolutional neural network (CNN) architecture tailored for mobile and embedded systems. It was proposed by Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun in 2017, aiming to strike a balance between computational efficiency and performance, making it well-suited for real-time applications on devices with limited processing power and energy restrictions.The primary innovation of ShuffleNet lies in its "channel shuffle" operations, which enhance the flow of information between feature channels in the network. To achieve this, the input feature channels are split into several groups, and the channels within and across these groups are then rearranged. This process enables more effective cross-group information sharing without substantially increasing the computational burden. In addition to channel shuffling, ShuffleNet incorporates depthwise separable convolutions and pointwise group convolutions to further minimize computational complexity. Depthwise separable convolution distinguishes between spatial and channel-wise convolutions, resulting in considerable computational savings compared to conventional convolution layers. Pointwise group convolutions, on the other hand, segregate channels into groups and apply convolutions individually within each group, further contributing to reduced computation.

Table 1. Measurement of accuracy and GPU speed of four different lightweight models with the same level of FLOPs on COCO object detection. (FLOPs: float-point operations.)

Model	mmAP (%)				GPU Speed (Images/s)			
FLOPs	40M	140M	300M	500M	40M	140M	300M	500M
Xception	21.9	29.0	31.3	32.9	178	131	101	83
ShuffleNet v1	20.9	27.0	29.9	32.9	152	85	76	60
MobileNet v2	20.7	24.4	30.0	30.6	146	111	94	72
ShuffleNet v2	23.7	29.6	32.2	34.2	183	138	105	83

Table 1 demonstrates that among the four distinct architectures, ShuffleNet V2 outperforms the others in terms of both accuracy and speed. As a result, we employ ShuffleNet V2 as the front-end component of our model. This architecture primarily incorporates two innovative operations: pointwise group convolution and channel shuffle. These operations significantly decrease the computational burden while maintaining the model's recognition accuracy.

The attention mechanism is Convolution Block Attention Mechanism (CBAM). It is an attention-based technique developed to boost the performance of convolutional neural networks (CNNs) by adaptively focusing on critical features in the input. Introduced in 2018 by Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, CBAM is a versatile and lightweight component that can be effortlessly incorporated into existing CNN structures without substantially impacting model complexity or computational demand. CBAM comprises two consecutive attention mechanisms: channel attention and spatial attention. Channel attention aims to emphasize relevant channels by leveraging the relationships between channels. This is achieved through global average pooling and global max pooling operations, followed by a common multi-layer perceptron (MLP). The generated channel-wise attention map is subsequently multiplied with the input feature map to accentuate crucial channels. After applying channel attention, the spatial attention mechanism concentrates on capturing spatial information by examining the relationships between spatial features within the feature map. It employs both average and max pooling operations over the channels, merges the results, and processes the combination through a convolutional layer with a sigmoid activation function. This process yields a spatial attention map, which is then element-wise multiplied with the channel-focused feature map.

The backend network is Gated Recurrent Unit (GRU). It is a recurrent neural network (RNN) architecture introduced in 2014. GRUs were designed to address the vanishing gradient problem often encountered in traditional RNNs, which hampers the network's ability to learn long-term dependencies in sequences. GRU has a simpler structure than the Long Short-Term Memory (LSTM) architecture, another widely-used RNN variant, while still offering comparable performance. GRU contains two gates: the reset gate and the update gate. These gates control the flow of information inside the hidden state, enabling the model to capture dependencies across varying time scales. The reset gate determines how much of the previous hidden state should be combined with the new input, allowing the model to decide the extent of the past information to be retained. The update gate, on the other hand, controls the extent to which the hidden state is updated with new information, balancing between maintaining the old hidden state and incorporating new insights.

EXPERIMENTAL SETTINGS AND RESULTS

In our study, we employ the OpenCV library to extract 29 frames from each video, focusing on the lip region by identifying 68 facial landmarks using the dlib library. We then resize all frames to 112x112 pixels, normalize them, and convert them to grayscale. To build and train our models, we utilize the open-source Tensorflow and Keras libraries, which offer user-friendly APIs.

Our model is trained on servers equipped with four NVIDIA Titan X GPUs. We divide the dataset into training and testing subsets at an 8:2 ratio, setting the epoch and batch size at 6032. The Adam optimizer is employed with an initial learning rate of 3×10^{-4}. Both the frontend and backend of our network are pre-trained on the LRW dataset. During training, we apply dropout with a 0.5 probability, and we measure our model's performance using the standard Cross Entropy loss.

In the Figure 2, as illustrated by the bar chart displaying accuracy and recall rates, the lowest recall rate for lip sign language recognition is associated with the number 0, followed by numbers 3 and 7. The accuracy metric for number 7 is markedly lower compared to the other nine digits. This is primarily attributed to the similarity in gestures between numbers 0 and 7. In the lip recognition system, numbers 0 and 7 exhibit low recall rates, while numbers 6, 7, and 8 have low accuracy rates. This is mainly due to the analogous

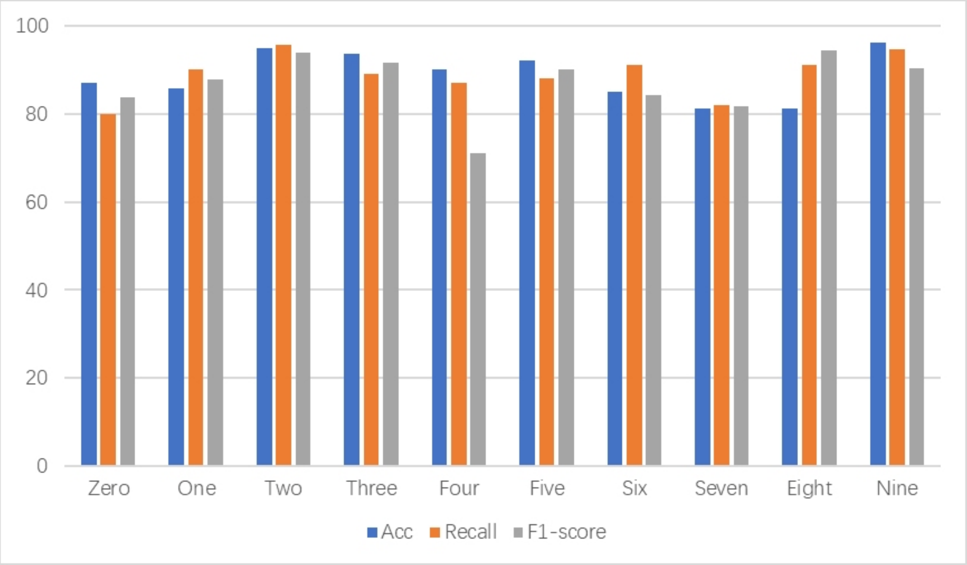

Figure 2: Prediction of pronunciation of English numbers.

lip movement patterns when pronouncing numbers 6, 7, and 8 in English. Consequently, to enhance their accuracy and recall rates, we should augment the training sample size for these specific numbers in the future.

CONCLUSION

Lip recognition technology offers significant potential and application demand in the realms of computer vision and human-computer interaction. Specifically, employing automatic lip recognition technology to enhance the social interactions of individuals with hearing and speech impairments is one of the most promising applications of artificial intelligence in healthcare and rehabilitation. Therefore, Our research primarily utilizes a combination of a lightweight network named ShuffleNet which dramatically reduces computational power for practical scenarios and GRU network. Additionally, we insert an attention mechanism into GRU to focus on more important information in feature maps. Furthermore, to validate the algorithm model's effectiveness, we appropriately apply a lab-generated test dataset for experimental testing. The result demonstrates that our model not only achieves comparable recognition accuracy but also reduces computational demands. In the future, we will do more research to improve the speed and accuracy of the model and expand self-built datasets for more tests.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (61971007 and 61571013).

REFERENCES

Assael, Y. M.; Shillingford, B.; Whiteson, S. Lipnet: End-to-end sentence-level lipreading. arXiv 2016, arXiv:1611.01599.

Bai, S.; Kolter, J. Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv 2018, arXiv:1803.01271.

He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

Hussein, D.; Ibrahim, D. M.; Sarhan, A. M. HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural Networks. Comput. Mater. Contin. 2021, 68, 1531–1549.

Lu, H.; Liu, X.; Yin, Y.; Chen, Z. A Patent Text Classification Model Based on Multivariate Neural Network Fusion. In Proceedings of the 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), Johannesburg, South Africa, 19–20 November 2019.

Mcgurk, H.; Macdonald, J. Hearing lips and seeing voices. Nature 1976, 264, 746–748.

Nazari, K.; Ebadi, M. J.; Berahmand, K. Diagnosis of alternaria disease and leafminer pest on tomato leaves using image processing techniques. J. Sci. Food Agric. 2022, 102, 6907–6920.

Palecek, K. Utilizing lipreading in large vocabulary continuous speech recognition. In Proceedings of the International Conference on Speech and Computer, Hatfield, UK, 12–16 September 2017; Springer: Berlin, Germany; Cham, Switzerland, 2017; pp. 767–776.

Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

Rostami, M.; Berahmand, K.; Nasiri, E.; Forouzandeh, S. Review of swarm intelligence-based feature selection methods. Eng. Appl. Artif. Intell. 2021, 100, 104210.

Saberi-Movahed, F.; Rostami, M.; Berahmand, K.; Karami, S.; Tiwari, P.; Oussalah, M.; Band, S. S. Dual Regularized Unsupervised Feature Selection Based on Matrix Factorization and Minimum Redundancy with application in gene selection. Knowl. Based Syst.2022, 256, 109884.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vincent, V.; Andrew, R. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.